Statistisches Matching: Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS

Bacher, Johann

Statistisches Matching: Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS

von Johann Bacher [1]

Zusammenfassung

Aufgabe des statistischen Matching ist das Auffinden von statistischen Zwillingen. Statistische Zwillinge sind dadurch gekennzeichnet, dass sie sich von ihren statistischen Zwillingsgeschwistern in ausgewählten Merkmalen nicht unterscheiden. Sie können für ein breites Spektrum von Aufgabenstellungen eingesetzt werden. In der sozialwissenschaftlichen Praxis ist ihre Anwendung – abgesehen von der Behandlung fehlender Werte – noch wenig verbreitet. Eine Ursache hierfür sind vermutlich fehlende Programmmodule in Standardstatistikprogrammen, wie SPSS. Das Hauptziel des Beitrages ist daher darzustellen, wie statistische Zwillinge mit Hilfe eines SPSS-Syntaxprogrammes berechnet werden können. Syntaxprogramme für zwei Methoden werden erörtert, nämlich für Propensity Scores und Distanzfunktionen. Das Vorgehen und die Berechnung werden anhand eines Forschungsbeispiels aus dem ALLBUS 1996 dargestellt.

Abstract

Statistical matching has the purpose of finding statistical twins. Statistical twins are cases that resemble their statistical siblings in selected variables. They can be applied to a lot of problems. However, they are – except for methods for imputing missing values – rarely used. Missing modules in standard statistical software are one reason for this situation. To describe how statistical twins can be computed with SPSS's Syntax is, therefore, one of the main aims of this paper. Two methods (propensity scores and distance functions) are discussed using the ALLBUS 1996 as an example.

1 Dr. ***Johann Bacher*** ist Professor am Lehrstuhl für Soziologie der Wirtschafts- und Sozialwissenschaftlichen Fakultät an der Universität Erlangen-Nürnberg, Findelgasse 7-9, D-90402 Nürnberg, Tel.: ++49-(0)911-5302-680, Fax.: ++49-(0)911-5302-660.

1. Definitionen und Anwendungsmöglichkeiten

Mitunter richtet sich das Forschungsinteresse auf die Wirkung einer Variablen, die nur bei wenigen Personen (bzw. allgemein Fällen) auftritt. Es soll beispielsweise mit Hilfe einer repräsentativen Umfrage, wie dem ALLBUS oder dem SOEP, untersucht werden, ob Arbeitslosigkeit ausländerfeindlich macht, ob der Aufstieg in eine Führungsposition zur Übernahme von konservativen geschlechtsspezifischen Rollenbildern führt, oder ob bei arbeitslosen Jugendlichen, die sich in einer Kursmaßnahme befinden, die Integration in den Arbeitsmarkt zunimmt. Um die Wirkung dieser Variablen (in unseren Beispielen Arbeitslosigkeit, Aufstieg in eine Führungsposition, Arbeitslosigkeit bei Jugendlichen) zu bestimmen, muss der Einfluss von anderen Variablen auf die abhängige Variable kontrolliert werden. Der übliche Weg über eine multivariate Analyse kann hierbei zu einer verzerrten Schätzung der Wirkung der untersuchten Variablen führen. Eine Möglichkeit, dieses schwerwiegende Problem zu überwinden, bietet das statistische Matching.

Formal lässt sich die Ausgangssituation des statistischen Matching wie folgt beschreiben: Es soll die Wirkung einer Variablen, die als B bezeichnet werden soll, auf eine andere Variable Y untersucht werden. . B hat zwei Ausprägungen, B_1 und B_2. B_1 tritt selten auf und die Personen mit der Merkmalsausprägung B_1 unterscheiden sich des Weiteren deutlich in weiteren Variablen X_i (i=1, 2, .., p) von den Personen aus B_2. Die Variablen X_i üben ebenfalls einen Einfluss auf Y aus. B könnte z.B. die Variable Arbeitslosigkeit mit den Ausprägungen ja (=B_1) und nein (=B_2) sein, X_1 das Alter, X_2 die Bildung, X_3 Gewerkschaftsmitgliedschaft usw. und Y die Ausländerfeindlichkeit (siehe dazu ausführlich Abschnitt 2). In dem Beispiel sind alle Ausgangsbedingungen erfüllt: Der Anteil der Arbeitslosen wird in einer Untersuchung im Allgemeinen im Vergleich zu Nicht-Arbeitslosen klein sein. Alter, Bildung und politische Selbsteinstufung korrelieren sowohl mit der Arbeitslosigkeit (Arbeitslose sind älter, höher gebildet und möglicherweise weniger häufig Gewerkschaftsmitglied) als auch mit der Ausländerfeindlichkeit (Ältere, Personen mit geringerer Bildung und fehlender Gewerkschaftsmitgliedschaft sind ausländerfeindlicher). Geschätzt werden soll der Effekt von B auf Y, also von Arbeitslosigkeit auf ausländerfeindliche Einstellungen. I.d.R. ist hierfür eine multivariate Analyse, z.B. eine lineare Regression oder ein Strukturgleichungsmodell, geeignet. Allerdings können dabei forschungslogische und forschungspraktische Probleme auftreten (siehe dazu ausführlich Abschnitt 2): B_2 (die Gruppe der "Nicht-Arbeitslosen") kann beispielsweise so *heterogen* sein, so dass sprichwörtlich Birnen mit Äpfeln verglichen werden mit der Folge, dass bei einer multivariaten Analyse verzerrte Schätzergebnisse auftreten. Eine Möglichkeit, eine besser definierte Vergleichsgruppe zu

ermitteln, bietet das statistische Matching: Im Idealfall sollen sich Untersuchungs- und Vergleichsgruppe nur in *B* unterscheiden, nicht aber in den Variablen X_i.

Beim statistischen Matching werden für jede Person *g* aus B_1 in B_2 ein oder mehrere Fälle *g** gesucht, die sich von der Person *g* in den Variablen X_i nicht oder nur geringfügig unterscheiden, also z.B. eine Person, die gleich alt ist, dieselbe Schulbildung aufweist und dasselbe Geschlecht hat wie der Fall *g*. Die Fälle *g** werden als *statistische Zwillinge* bezeichnet. Für die Suche nach statistischen Zwillingen wird die Bezeichnung *statistisches Matching*[2] (Diese und die folgenden Fußnoten sind nicht hochgestellt!) oder einfach nur *Matching* verwendet.

Statistische Zwillinge eignen sich aber nicht nur zur Schätzung der Wirkung einer Untersuchungsvariablen. Sie können auch für folgende, derzeit "modernere" Aufgabenstellungen eingesetzt werden:

- *Schätzung bzw. Imputation von fehlenden Werten.* Der Datensatz g hat in einer Variablen *Y* (z.B. Einkommen oder Parteipräferenz) einen fehlenden Wert. Zur Schätzung dieses fehlenden Wertes wird ein statistischer Zwilling von *g* gesucht, der in den Variablen X_i (z.B. Geschlecht, Alter, Bildung, sozialer Status, berufliche Position usw.) ähnliche Merkmalsausprägungen besitzt.[3]
- *Datenfusion.* Zwei Datenfiles (z.B. eine Untersuchung über den Medienkonsum und eine über das Freizeitverhalten) sollen über eine Menge gemeinsamer Merkmale X_i (z.B. Alter, Einkommen, Bildung, Haushaltsform, Freizeitinteressen) fusioniert werden. Zu jedem Datensatz *g* aus dem ersten Datenfile (Empfängerdatei, z.B. aus der Untersuchung über den Medienkonsum) wird ein statistischer Zwilling *g** aus dem zweiten Datenfile (Spenderdatei, z.B. dem Freizeitverhalten) mit ähnlichen Ausprägungen in den Merkmalen X_i gesucht. Ziel ist u.a. die Ermittlung eines Zusammenhangs zwischen Variablen, die jeweils nur in einer der beiden Befragungen erfasst wurden. In der Medienkonsumbefragung wurde z.B. erhoben, ob die Berichterstattung über die Tour de France im Fernsehen verfolgt wurde, in der zeitlich parallel durchgeführten Studie über das Freizeitverhalten wurde gefragt, ob eine Ferienreise nach Frankreich geplant sei.[4]

---

2 Davon zu unterscheiden ist die Optimal Matching Analyse oder Sequenzanalyse (***Aisenbrey*** 2000, S. 13-34), bei der Ähnlichkeiten zwischen Verlaufsmustern, z.B. Lebensläufen, gemessen und untersucht werden.

3 Konkrete Anwendungen werden sehr ausführlich und anschaulich dargestellt in ***Holm*** (2001, S. 57-90).

4 Siehe dazu ebenfalls ***Holm*** (2001, S. 92-126). Anwendungen aus der Medienforschung werden dargestellt in ***Rässler*** (2001), der wohl derzeit umfassendsten und systematischsten Behandlung der Datenfusion.

Hinzukommt die bereits behandelte *Bestimmung einer Kontrollgruppe:* Zu einer Untersuchungsgruppe (z.B. eine Gruppe von arbeitslosen Jugendlichen, die an einer Kursmaßnahme teilgenommen haben) soll zur Effektschätzung eine Kontrollgruppe aus einem bekannten Register (z.B. dem Arbeitslosenregister) oder einer vorhandenen Untersuchung (z.B. ALLBUS, SOEP) gezogen werden, die sich hinsichtlich einer Menge von Kontrollvariablen X_i (z.B. Geschlecht, Alter, schulische und berufliche Vorbildung, Wohnort usw.) nicht von der Untersuchungsgruppe unterscheidet.[5]

Für statistische Zwillinge gibt es somit ein breites und interessantes Feld von Anwendungen, das auch theoretisch in der einschlägigen Fachliteratur intensiv diskutiert wurde und wird. Die Arbeiten von Rubin und Little (***Rubin*** 1987; ***Little*** und ***Rubin*** 1987) zur Behandlung von fehlenden Werten gelten bereits ebenso als Klassiker wie Beiträge von Heckman u.a. (zit. z.B. in ***Lechner*** 1999) und von Rubin, Rosenbaum oder Holland (zit. z.B. in ***Lechner*** 1999 und ***Smith*** 1997) auf dem Gebiet der statistischen Effektschätzung. Schließlich wurde 2001 von ***Rässler*** (2001) auch ein umfassender und systematischer Überblick über die Datenfusion vorgelegt.

Als weitere Möglichkeit der Anwendung des Matching kann noch die *Berechnung des empirischen Re-Identifikationsrisikos* von Registerdaten angeführt werden. Dadurch kann untersucht werden, ob eine Weitergabe von Registerdaten von öffentlichen Behörden, wie der Bundesanstalt für Arbeit oder dem statistischen Bundesamt, an Private möglich ist, da eine Re-Identifikation von einzelnen Fällen faktisch ausgeschlossen werden kann. Zur Abschätzung des Re-Identifikationsrisikos könnte wie folgt verfahren werden: Zu jedem Datensatz des Anwenders wird ein statistischer Zwilling in den registrierten Daten gesucht. Anschließend wird geprüft, ob der statistische Zwilling der reale Zwilling ist.[6] Überschreitet der Prozentsatz der richtig gefundenen Zwillinge einen bestimmten Schwellenwert, werden die Daten vor der Weitergabe weiter anonymisiert.[7]

---

5 Siehe dazu die Arbeiten von Lechner u.a. zur Evaluation der aktiven Arbeitsmarktpolitik (***Lechner*** 1999; ***Gerfin*** und ***Lechner*** 2000; ***Fröhlich*** 2002)

6 Eine empirische Analyse von Bender u.a. erbrachte für die Beschäftigtenstichprobe des IAB (Registerdaten) und eine MPI-Umfrage ein Risiko von etwa 9%. (***Bender***, ***Brand*** und ***Bacher*** 2001; ***Bacher***, ***Brand*** und ***Bender*** 2002).

7 Diese Möglichkeit wird derzeit nicht genutzt. Die statistischen Behörden haben andere Wege eingeschlagen, wie die Überlagerung der Daten mit Zufallsfehlern oder die Aggregierung (Zusammenfassung) von Merkmalsausprägungen. (***Brand*** 2000 sowie die Beiträge in ***United Nations*** 2001)

Bedauerlichweise bieten Standardstatistikprogramme, wie SPSS – abgesehen von rudimentären Imputationstechniken für fehlende Werte – keine Programmmodule für die genannten Einsatzmöglichkeiten an. Daher soll nachfolgend dargestellt werden, wie mit Hilfe eines Syntaxprogramms in SPSS statistische Zwillinge ermittelt werden können.

2. Anwendungsbeispiel

Zur Verdeutlichung des Vorgehens soll mit Hilfe des ALLBUS 1996 (ZA-Studiennummer 2800) die Frage untersucht werden, ob Arbeitslosigkeit zu Ausländerfeindlichkeit führt. Für die Analyse wurden aus dem ALLBUS 1996 in einem ersten Schritt Personen ausgewählt, die in Deutschland geboren wurden und die deutsche Staatsbürgerschaft besitzen , da unklar ist, wie Ausländer Fragen zur Ausländerfeindlichkeit beantworten: Meinen sie mit "Ausländern" oder "Fremden" sich selbst oder Deutsche?[8] In einem zweiten Schritt wurden Personen selektiert, die zum Befragungszeitpunkt erwerbstätig (ganztags oder halbtags) oder arbeitslos waren. Es verblieben n=1855 Personen[9], von denen zum Befragungszeitpunkt 187 (10,1%) arbeitslos waren.

Für die Analyse wurde die Annahme getroffen, dass die Ausländerfeindlichkeit neben dem Erwerbsstatus von folgenden weiteren Variablen abhängt:

- Erhebungsgebiet (V3)
- Alter (V37)
- Geschlecht (V141)
- Familienstand (V183)
- allgemeiner Schulabschluss (V142)
- Berufsprestige (V160 bei Erwerbstätigen und V176 bei Arbeitslosen)
- Konfession (V318)
- Kirchgangshäufigkeit (V319)
- politischen Selbsteinstufung (V112)
- Gewerkschaftsmitgliedschaft (V320)

---

8 Empirisch ergeben sich signifikante Unterschiede in fünf der sieben untersuchten Items zur Ausländerfeindlichkeit zwischen beiden Gruppen (in Deutschland geborene Personen mit deutscher Staatsbürgerschaft versus andere Personen). Unsere Untersuchungsgruppe (in Deutschland geborene Personen mit deutscher Staatsbürgerschaft) äußert sich in den fünf Items ausländerfeindlicher.

9 Wegen fehlender Werte mussten in der Folge weitere Personen eliminiert werden, siehe dazu unten. Einbezogen werden konnten schließlich 1809 Personen.

Als abhängige Variablen wurden diskriminierende und integrierende Aussagen über Ausländer (V72 bis V78), wie z.B. "Ausländer belasten unser soziales Netz" (V73) oder "Ausländer stützen die Rentenversicherung" (V76) in die Analyse einbezogen. Die Items bilden eine eindimensionale Skala (1. Eigenwert = 2,78; 2. Eigenwert = 0,97; Ergebnisse einer Hauptkomponentenanalyse), wenn die Variable V72 ("Ausländer tun die unschönen Arbeiten") aus der Analyse ausgeschlossen wird[10]. Zur Überprüfung der Quantifizierbarkeit der ordinalen Kategorien wurde eine multiple Korrespondenzanalyse (MCA) gerechnet (***Blasius*** und ***Thiessen*** 2001; ***Blasius*** 2001, S. 338-346). Die MCA bestätigte den Befund der Hauptkomponentenmethode. Es ergibt sich das von Guttman bereits in den 50er Jahren gefundene typische Muster für eindimensionale Skalen (***Bacher*** 1996, S. 127; ***Blasius*** 2001, S. 343). Die erste Dimension misst die Zieldimension und bildet die Schwierigkeitsgrade ab. Alle Antwortkategorien sind geordnet, d.h. die Antwortkategorie 1 hat einen kleineren (oder größeren) Skalenwert auf der ersten Dimension als die Kategorie 2, diese hat wiederum einen kleineren (oder größeren) Skalenwert als die Kategorie 3, diese wiederum einen kleineren (oder größeren) als die Kategorie 4 usw.

In die weitere Analyse gingen nicht die Einzelitems ein, sondern der erste Faktor der durchgeführten Hauptkomponentenanalyse[11].

Das übliche Vorgehen zur Ermittlung des Einflusses der Arbeitslosigkeit auf die Ausländerfeindlichkeit würde nun darin bestehen, dass eine multiple Regression mit der Ausländerfeindlichkeit als abhängiger Variablen durchgeführt wird. Als unabhängige Variable würden neben dem Erwerbstatus die oben angeführten Kontrollvariablen (Erhebungsgebiet, Geschlecht usw.) einbezogen werden.

Dieses Vorgehen ist nicht unproblematisch. Folgende forschungslogische und forschungspraktische Bedenken (***Chapin*** 1974, S. 34[12]; ***Rässler*** 2001, S. 24; ***Smith*** 1997, S. 326-327) lassen sich anführen:

---

10 Offensichtlich fällt bereits ausländerfeindlich orientierten Personen eine Verneinung dieser Aussage schwer.

11 An Stelle des Faktorwertes hätte – worauf der Gutachter dieses Beitrages zu Recht hinweist – ein inhaltlich besser interpretierbarer Gesamtpunktwert verwendet werden können. Dieser korreliert mit 0,998 mit dem Faktorwert aus der Hauptkomponentenmethode bzw. mit 0,990 mit den Personenscores auf der ersten Dimension der multiplen Korrespondenzanalyse, sodass es für eine Zusammenhangsanalyse keine Rolle spielt, welche der drei Variablen (Gesamtpunktwerte, Faktorwerte, Personenscores) verwendet wird.

12 ***Chapin*** (1974) führt in seiner heute weitgehend in Vergessenheit geratenen Abhandlung über experimentelle Designs in der Soziologie folgende Aspekte an: "It is, however, not advisable to use partial correlation unless certain conditions of reliable measurement, homogeneity,

- *Logisch unzulässiger Vergleich:* Es werden zwei nicht vergleichbare Gruppen gegenübergestellt. Der Vergleichsgruppe gehören sehr unterschiedliche Personen an. Einige von ihnen haben kein oder nur ein sehr geringes Arbeitslosigkeitsrisiko. Besser wäre es eine Vergleichsgruppe zu bilden, die dasselbe oder ein vergleichbares Arbeitslosenrisiko besitzt wie die Untersuchungsgruppe. Dadurch würde man sich besser experimentellen Versuchsbedingungen annähern und der Effekt könnte unverzerrter geschätzt werden.
- *Dominanz der nicht relevanten Vergleichsgruppe:* Die Ergebnisse hängen primär von der relativ großen Vergleichsgruppe ab.
- *Heterogenität der nicht relevanten Vergleichspopulation:* Die Vergleichsgruppe ist i.d.R. sehr groß und daher heterogen. Sie setzt sich möglicherweise aus unterschiedlichen Teilpopulationen mit unterschiedlichen Wirkungszusammenhängen zusammen.

Tabelle 1 verdeutlicht die Auswirkungen einer heterogenen Vergleichsgruppe. Es wurde von folgenden Annahmen ausgegangen. Gegeben sind zwei Gruppen B_1 und B_2. Beide Gruppen sind mit je 100 Fällen gleich groß (n_1=n_2=100) und unterscheiden sich nicht in X, d.h. $\rho_{XB} = 0$ (theoretische Vorgabe) bzw. nahe bei Null (Simulationsergebnis $r_{XB} = 0{,}020$). Die Variable B korreliert dagegen mit Y mit einer bestimmten Stärke, in dem Beispiel mit $r_{YB} = 0{,}155$ (Simulationsergebnis), während die Korrelation zwischen ρ_{YX} gleich 0 (theoretische Vorgabe) bzw. nahe bei Null (Simulationsergebnis $r_{YX} = 0{,}016$) ist, d.h. die Kontrollvariable hat keinen Einfluss auf Y. Die partielle Korrelation $r_{YB/X}$ zwischen B und Y gegeben X ist daher nicht von der bivariaten Korrelation ($r_{YB} = 0{,}155$) verschieden und hat den Wert 0,155. Sie ist statistisch signifikant von Null verschieden. B hat somit einen signifikanten Einfluss auf Y.

In der weiteren Folge wurde angenommen, dass die "echte" Vergleichsstichprobe B_2 nicht verfügbar ist, sondern die Population B_{2*}, die sich aus B_2 und einer weiteren Subpopulation B_3 zusammensetzt, in die Analyse einbezogen werden soll. Im Unterschied zu B_1 und B_2 besteht in B_3 ein Zusammenhang zwischen X und Y. Die theoretische Korrelation ρ zwischen X und Y innerhalb von B_3 beträgt 0,500. B_3 unterscheidet sich ferner von B_2 und B_1 dadurch, dass in X höhere Variablenwerte auftreten (Mittelwert in X ist gleich 4 im Unterschied zu 2). Ferner ist B_3 mit $n_3 = 800$ deutlich größer und hat in X eine höhere Streuung (an Stelle von 0,5 beträgt die Standardabweichung 1,0). Die Verwendung von $B_{2*} = \{B_2, B_3\}$ an

sampling, and normality of distribution are met." (S. 34) Leider erörtert er diese Aspekte nicht weiter.

Stelle von B_2 führt dazu, dass die Variablen *B* und *X* einerseits und *B* und *Y* andererseits stark negativ miteinander korrelieren, während zwischen *X* und *Y* eine relativ starke positive Korrelation berechnet wird. Unter diesen Modellannahmen ergibt sich eine partielle Korrelation $r_{YB/X}$ von 0,018. Der Wert weicht nicht statistisch signifikant von Null ab. Der ursprünglich positive Zusammenhang von 0,155 verschwindet somit.

Tabelle 1: Auswirkungen einer heterogenen Vergleichsgruppe auf die Korrelation der Behandlungsvariablen B mit der abhängigen Variablen Y

	Spezifikation	Simulationsergebnisse
A	• Untersuchungsgruppe B_1: 100 Fälle verteilen sich auf X normal mit Mittelwert 2 und Standardabweichung 0,5; X ist unkorreliert mit Y in B_1. • Vergleichsgruppe B_2: 100 Fälle verteilen sich auf X normal mit Mittelwert 2 und Standardabweichung 0,5; X ist unkorreliert mit Y in B_2.	bivariate Korrelationen B 1,000 X 0,020 1,000 Y 0,155 0,016 1,000 $r_{YB/X}$ = 0,155 (p<5%)
B	• Untersuchungsgruppe B_1: Spezifikation siehe oben • Vergleichsgruppe B_{2*}: setzt sich zusammen aus B_2 und B_3; - Spezifikation von B_2 siehe oben; - Spezifikation von B_3: 800 Fälle, die sich auf X normal verteilen mit Mittelwert 4 und Standardabweichung 1; X korreliert mit Y mit ρ = 0,500 in B_3, in den anderen Gruppen (B_1 und B_2) ist die Korrelation 0	bivariate Korrelationen B 1,000 X -0,447 1,000 Y -0,208 0,497 1,00 $r_{YB/X}$ = 0,018 (p>10%)
C	wie Situation B, nur dass X und Y in B_3 mit ρ = 0,800 positiv korrelieren	$r_{YB/X}$ = -0,075 (p<5%)
D	wie Situation B, nur dass X und Y in B_3 mit ρ = -0,500 positiv korrelieren	$r_{YB/X}$ = 0,294 (p<0,1%)

$r_{BY/X}$=partielle Korrelation zwischen B und Y bei Kontrolle von X

Denkbar ist aber auch, dass sich die Richtung des Zusammenhangs umkehrt (Situation C in Tabelle 1: es tritt eine – zwar numerisch kleine – signifikante negative partielle Korrelation auf) oder der ursprüngliche Zusammenhang verstärkt wird (Situation D in Tabelle 1). Die Verzerrung (Verstärkung, Abschwächung, Umkehrung) hängt von der Größe der zusätzlichen Subpopulation B''_2 ab, von deren Verteilung auf X und der Korrelation zwischen X und Y.

Eine heterogene Vergleichsgruppe kann auch dazu führen, dass ein Zusammenhang erzeugt wird, der nicht besteht. Ab welchem Grad an Heterogenität Verzerrungen auftreten und welcher Faktor (ungleiche Stichprobengröße, ungleiche Lage auf X, unterschiedliche Korrelationen in Subpopulationen) welche Bedeutung hat, lässt sich allgemein schwer bestimmen.

Das Beispiel sollte aber verdeutlichen, dass es beim Vorliegen einer heterogenen Vergleichsgruppe vorteilhaft sein kann, eine homogenere Vergleichsgruppe zu bilden, um Verzerrungen der dargestellten Art zu vermeiden. Wie dies praktisch mit Hilfe des statistischen Matching gemacht werden kann, soll nachfolgend dargestellt werden.

Anzumerken ist, dass die hier in einer Simulationsstudie untersuchte Konstellation empirisch durchaus auftritt. Arbeitslosigkeit konzentriert sich z.B. auf untere Bildungsstufen. Betrachtet man nur diese, ergibt sich in Westdeutschland kein signifikanter Zusammenhang von $r = 0{,}050$ ($p>10\%$) zwischen Arbeitslosigkeit und Ausländerfeindlichkeit. Werden dagegen alle Bildungsstufen einbezogen, nimmt man also die Gruppe der höher gebildeten, i.d.R. erwerbstätigen Personen ($=B_3$ in unserer Simulationsrechnung) hinzu, erhöht sich die Korrelation auf $r = 0{,}074$ ($p<5\%$) und wird statistisch signifikant.

3. Verfahren zur Suche von statistischen Zwillingen (Matching-Verfahren)

Im Folgenden soll zur Beschreibung des Verfahrens davon ausgegangen werden, dass zwei Datenfiles vorliegen, eine Empfängerdatei (in unserem Beispiel die Arbeitslosen), die alle Fälle mit dem untersuchten Ereignis umfasst, und eine Spenderdatei (in unserem Beispiel die Erwerbstätigen). Zu jedem Datensatz g des Empfängerfiles soll im Spenderfile ein statistischer Zwilling g* gesucht werden. Dafür sind drei Entscheidungen erforderlich:

- Auswahl von geeigneten Variablen,
- Auswahl eines Suchverfahrens und
- Auswahl eines Verfahrens zur Berechnung der Ähnlichkeit.

Auswahl der Variablen. In der Literatur wird empfohlen, alle Variablen in die Berechnung der Propensity Scores einzubeziehen, die einen Einfluss auf die abhängige Variable(n) haben können. (***Rubin*** und ***Thomas*** 1996 zit. in ***Smith*** 1997: 335). In unserem Anwendungsbeispiel sind dies alle oben angeführten unabhängigen Variablen (Erhebungsgebiet, Alter, Geschlecht, Familienstand usw.). Alternativ wurde vorgeschlagen, nur Schlüsselvariablen, die einen signifikanten Einfluss auf die abhängige Variable haben, zu verwenden (***Smith*** 1997, S. 338). Diese Schlüsselvariablen sind aber a priori meistens nicht bekannt, so dass dieser Vorschlag häufig nicht anwendbar ist.

Auswahl eines Suchverfahrens. Ein einfaches Verfahren besteht darin, dass die beiden Datenfiles in eine zufällige Anordnung gebracht werden. Daran anschließend wird mit dem ersten Datensatz in der Empfängerdatei begonnen. Für diesen Datensatz wird ein statistischer Zwilling in der Spenderdatei gesucht, beginnend mit dem ersten Datensatz. Kommen mehrere Datensätze als statistische Zwillinge in Frage, wird der erste ausgewählt. Für die weitere Suche wird der statistische Zwilling gestrichen. Die Suche wird für den zweiten Datensatz der Empfängerdatei fortgesetzt usw. Der beschriebene Suchalgorithmus mit der Bezeichnung "random order, nearest available pair-matching method" (***Smith*** 1997, S. 338) lässt sich wie folgt formalisieren:

1. Ordne die Fälle in der Empfänger- und Spenderdatei zufällig an.
2. Setze g in der Empfängerdatei gleich 1.
3. Suche für g in der Spenderdatei einen statistischen Zwilling g*. Beginne dafür mit dem ersten Datensatz. Kommen mehrere g* als statistische Zwillinge in Frage, wähle g* mit dem kleinsten Index.
4. Streiche g* für die weitere Suche.
5. Erhöhe g um 1 und wiederhole die Schritte 3 und 4 solange, bis alle Fälle g der Empfängerdatei abgearbeitet sind.

In der Regel wird zusätzlich gefordert, dass die Unterschiede zwischen g und g* kleiner einem bestimmten Schwellenwert sein müssen, damit g* als statistischer Zwilling betrachtet wird.

Auswahl eines Verfahrens zur Berechnung der Ähnlichkeit. Die Suche nach statistischen Zwillingen setzt eine Messung der Ähnlichkeit zwischen zwei Datensätzen voraus. Hierzu lassen sich zwei Vorgehensweisen unterscheiden: die Methode der so genannten Propensity Scores und Distanzmethoden. Den Propensity Scores wird heute üblicherweise der Vorzug gegeben. Sie sollen daher zuerst behandelt werden.

Propensity Scores. Bei der Methode der Propensity Scores wird zunächst eine logistische Regression gerechnet. Als abhängige Variable geht die untersuchte Gruppierungsvariable (Behandlungsvariable) ein, in unserem Beispiel ist dies der Erwerbstatus mit den Ausprägungen 1 für arbeitslos und 0 für erwerbstätig. [13] Als unabhängige Variablen werden alle ausgewählten Kontrollvariablen einbezogen: Erhebungsgebiet (V3), Alter (V37), Geschlecht (V141), Familienstand (V183), allgemeiner Schulabschluss (V142), Berufsprestige (V160 bei Erwerbstätigen und V176 bei Arbeitslosen), Konfession (V318), Kirchgangshäufigkeit (V319), politischen Selbsteinstufung (V112) und Gewerkschaftsmitgliedschaft (V320). Der Familienstand und die Konfession wurden für die Analyse in Dummies aufgelöst, da sie nominalskaliert sind. Die ordinalskalierte Variable Schulbildung wurde wie eine quantitative behandelt. Die mit nur wenigen Fällen besetzte Ausprägung "6" (anderer Schulabschluss) wurde als MISSING definiert.

Tabelle 2: Ergebnisse der logistischen Regression zur Berechnung von Propensity Scores

	Regressions-koeffizient b	Standardfehler	Fehlerniveau p
OST	1,029	,215	,000
ALTER	,039	,009	,000
WEIBL	,465	,181	,010
VERHGE	1,278	,501	,011
VERW	,326	,486	,502
GESCHIED	,714	,286	,012
LEDIG	,455	,253	,072
SCHULB	-,681	,137	,000
PRES	-,022	,005	,000
RK	-1,227	,333	,000
EV	-,441	,218	,043
ANDERE	1,497	,634	,018
LINKSRE	-,072	,050	,153
GEWERK	-,265	,220	,228
Konstante	-1,004	,682	,141

13 Verallgemeinerungen auf eine Behandlungsvariable mit mehr als zwei Ausprägungen werden erörtert in ***Lechner*** (1999) und ***Fröhlich*** (2002).

Die Ergebnisse der logistischen Regression fasst Tabelle 2 zusammen. Den Variablen wurden inhaltlich aussagekräftigere Namen gegeben. OST steht für Erhebungsgebiet und hat die Werte 1 (für Ostdeutschland) und 0 (für Westdeutschland). Für das Alter (V37) wurde der Variablenname ALTER verwendet usw. In die Analyse wurden alle Fälle der Empfänger- und Spenderdatei einbezogen.

Auf der Grundlage der Schätzergebnisse lässt sich für jede Person der Empfänger- und Spenderdatei ein Prognosewert (=Propensity Score) berechnen:

$$P(Erwerbsstatus = 1(\text{"}arbeitslos\text{"})) = P(Behandlung = 1) = \frac{e^z}{1+e^z} \text{ mit}$$

$$z = 1{,}029 \cdot OST + 0{,}039 \cdot ALTER + 0{,}465 \cdot WEIBL + \ldots + (-0{,}265) \cdot GEWERK + (-1{,}004)$$

Für jeden arbeitslosen Befragten g aus der Empfängerdatei wird nun unter den Erwerbstätigen der Spenderdatei ein statistischer Zwilling g* gesucht, der folgende Bedingungen erfüllt:

- $d(P_g, P_{g*}) = \left|P_g - P_{g*}\right| \rightarrow \min(d(P_g, P_{g'}))$ mit $P_g = P(Erwerbstatus\ von\ g = 1)$ und $P_{g*} = P(Erwerbstatus\ von\ g* = 1)$. Unter allen noch verfügbaren Fällen g' (also unter allen Fällen, die bisher noch nicht als Zwilling verwendet wurden) in der Spenderdatei (also der Datei der Erwerbstätigen bzw. allgemein der potentiellen Zwillinge) hat g* die kleinste Distanz (Abweichung) zu g.
- $d(P_g, P_{g*}) < c$. Die Distanz muss kleiner einem bestimmten Schwellenwert sein (z.B. kleiner 0,01; 0,001; 0,0001 usw.)

Für die Wahl des Schwellenwertes *c* gibt es keine allgemeinen Regeln. Wird der Schwellenwert *c* zu groß gewählt, besteht die Gefahr, dass "schlechte" Zwillinge gefunden werden, die sich in den Ausprägungen der Variablen deutlich unterscheiden. Wird ein zu kleiner Schwellenwert *c* festgelegt, können viele Fälle ohne Zwillinge bleiben. Es empfiehlt sich daher, unterschiedliche Werte von *c* beginnend mit einem sehr kleinen (z.B. *c*=0,0001) oder einem relativ großen Wert (*c*=0,2) auszuprobieren.

Wird in unserem Beispiel der Schwellenwert gleich 0,1 gesetzt, ergibt sich folgendes Zwillingspaar:

	V2	BEHANDL	OST	ALTER	WEIBL	VERH	VERHGE	VERW
Person g:	80	1,00	1,00	58,00	,00	1,00	,00	,00
stat. Zwilling:	927	,00	1,00	55,00	1,00	1,00	,00	,00

	GESCHIED	LEDIG	SCHULB	PRES	RK	EV	ANDERE
Person g:	,00	,00	2,00	50,70	,00	1,00	,00
stat. Zwilling:	,00	,00	2,00	85,40	,00	,00	,00

	KEINE	LINKSRE	GEWERK	PRE_1
Person g:	,00	5,00	,00	,27925
stat. Zwilling:	1,00	2,00	1,00	,27632

Lesehilfe: Die Person mit der Identifikationsnummer 80 (V2=80) gehört der Untersuchungsgruppe (BEHANDL = 1 = arbeitslos) an. Sie kommt aus dem Osten (OST = 1), ist 58 Jahre alt, männlich, verheiratet, hat eine geringe Bildung (2=Volks- bzw. Hauptschulabschluss), das Berufsprestige beträgt 50,70. Die Person ist evangelisch, befindet sich in der politischen Mitte und ist bei keiner Gewerkschaft. Ihr statistischer Zwilling mit der Identifikationsnummer 927 lebt ebenfalls im Osten, ist 55 Jahre, weiblich und verheiratet. Die Schulbildung ist gering, das Berufsprestige hat einen Wert von 85,40. Die Person ist konfessionslos, politisch links und Gewerkschaftsmitglied.

Die beiden Datensätze unterscheiden sich hinsichtlich des Alters, des Geschlechts, der Konfession, des Berufsprestiges, der Links-Rechts-Einstufung und der Gewerkschaftsmitgliedschaft. Der Altersunterschied ist mit 58 Jahren im Vergleich mit 55 Jahren relativ gering. Die Abweichungen im Berufsprestige fallen dagegen größer aus (50,70 versus 85,40), wobei aber zu beachten ist, dass der Regressionskoeffizient *b* für das Berufsprestige mit –0,022 relativ gering. Erhebungsgebiet, Familienstand und Bildung sind identisch.

Obwohl es sich hier nur um einen Einzelfall handelt, legt das Beispiel nahe, dass vielleicht ein strengerer Schwellenwert *c* (z.B. 0,001) als im vorliegenden Fall (*c*=0,1) gewählt werden sollte. Ein Vermutung, die weiterführende Analysen bestätigen werden.

Distanzmaße. Zur Ermittlung der Ähnlichkeit zwischen zwei Datensätzen wird keine neue abgeleitete Variable (die Prognosewerte oder Propensity Scores) erzeugt, sondern es werden Distanzen berechnet. Da die Variablen gemischtes Messniveau und unterschiedliche Skaleneinheiten haben, müssen sie transformiert bzw. gewichtet werden, um Vergleichbarkeit zu erzielen (***Bacher*** 1996, S. 173-198). Andernfalls würden die Ergebnisse fast ausschließlich von den Variablen mit der größten Variationsbreite bestimmt werden.[14]

Sollen (gewichtete) quadrierte euklidische oder euklidische Distanzen[15,16] verwendet werden, können folgende Gewichte definiert werden:

$$w_{ik} = \frac{1}{\sqrt{2} \cdot s_{ik}} = \frac{1}{\sqrt{2} \cdot \sqrt{p_{ik} \cdot (1 - p_{ik})}}$$ für die Dummies der nominalen Variablen

und

$$w_j = \frac{1}{s_j}$$ für die quantitativen Variablen,

wobei s_{ik} die Standardabweichung der k-ten Dummy der nominalen Variablen i ist. s_j ist die Standardabweichung der quantitativen Variablen j.

Im Unterschied zur logistischen Regression müssen alle Dummies einer nominalen Variablen in die Berechnung eingehen. Ordinale Variablen werden wie quantitative behandelt.[17] Bei nominalen Variablen müssen die Dummies zusätzlich mit dem

14 In unserem Beispiel wären dies die Variablen Alter und Prestige.

15 In der Literatur (z.B. ***Smith*** 1997) wird häufig auf die Mahalanobis-Distanz verwiesen. Die Mahalanobis-Distanz führt automatisch eine Gewichtung der Variablen durch, bei der ungleiche Varianzen und Korrelationen zwischen den Variablen beseitigt werden. Variablen, die stark miteinander korrelieren, erhalten ein geringeres Gewicht. Für nominale und damit gemischte Variable ist die Mahalanobis-Distanz nicht berechenbar, da sie eine Inversion der Varianz-Kovarianz-Matrix der Variablen voraussetzt. Diese mathematische Operation ist bei Verwendung aller Dummies (siehe dazu Textteil) einer nominalen Variablen nicht möglich. An Stelle der quadrierten euklidischen Distanz kann auch die euklidische Distanz verwendet werden.

16 An Stelle der quadrierten euklidischen Distanz kann auch die euklidische Distanz verwendet werden. Auf die Ergebnisse hat dies keinen Einfluss. Beide Distanzmaße sind gewichtete Koeffizienten, es wird also mit einer gewichteten (quadrierten) euklidischen Distanz gerechnet. Aus Gründen der sprachlichen Einfachheit wird nur von (quadrierten) euklidischen Distanzen gesprochen. Alle Berechnungsformeln beziehen sich auf die quadrierte euklidische Distanz.

17 Das formal richtigere Vorgehen wäre - da die für ordinale Variablen entwickelten Distanzmaße eine implizite Gewichtung beinhalten (***Bacher*** 1996, S. 216-218) und daher hier nicht geeignet sind - eine Behandlung als nominalskaliert. Dadurch entsteht aber ein beträchtlicher Informationsverlust. So z.B. würde sich in der Links-Rechts-Orientierung dieselbe Distanz ergeben, falls sich zwei Personen extrem unterscheiden (eine Person hat z.B. den Skalenwert 1, die andere den Skalenwert 10) und falls die Unterschiede nur sehr gering sind (eine Person hat z.B. den Skalenwert 1, die andere den Wert 2). Werden ordinale Variablen dagegen wie quantitative Be-

Kehrwert der Wurzel aus 2 gewichtet wird. Die Notwendigkeit dieser zusätzlichen Gewichtung lässt sich am Beispiel von dichotomen Variablen leicht verdeutlichen.[18] Dichotome Variablen lassen sich unmittelbar wie quantitative Variablen behandeln. Eine Auflösung in Dummies ist daher nicht erforderlich. Wird eine dichotome Variable direkt als quantitativ betrachtet, ist die maximale quadrierte euklidische Distanz zwischen zwei Datensätzen gleich 1. Selbstverständlich ist es auch möglich, eine dichotome Variable wie eine nominale Variable zu betrachten und in ihre zwei Dummies aufzulösen. Ohne Gewichtung wäre die maximale quadrierte euklidische Distanz dann allerdings gleich 2. Eine zusätzliche Gewichtung mit dem Kehrwert der Wurzel aus 2 löst dieses Problem.

Werden die Variablen gewichtet mit

$$a^*_{ik} = w_{ik} \cdot a_{ik} = \frac{a_{ik}}{\sqrt{2} \cdot s_{ik}} = \frac{a_{ik}}{s_{ik}} \cdot 0.7071$$

bzw. mit

$$x^*_j = w_j \cdot x_j = \frac{x_j}{s_j},$$

lässt sich die Distanz zwischen zwei Datensätzen g und g' berechnen als

$$d^2_{g,g'} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} (a^*_{gik} - a^*_{g'ik})^2 + \sum_{j=1}^{J} (x^*_{gj} - x^*_{g'j})^2 = \sum_{i=1}^{I} \sum_{k=1}^{K_i} w^2_{ik} \cdot (a_{gik} - a_{g'ik})^2 + \sum_{j=1}^{J} w^2_j (x_{gj} - x_{g'j})^2,$$

wobei a_{ik} die Dummies der nominalen Variablen i sind. Ein statistischer Zwilling g* von g ist dadurch gekennzeichnet, dass gilt:

- $d^2_{g,g^*} = \min(d^2_{g,g'})$ mit g' = noch verfügbares Element aus der Spenderdatei.
- $d^2_{g,g^*} < c$.

Die Wahl des Schwellenwertes ist etwas schwieriger als jene bei der logistischen Regression, da das Maximum der (quadrierten) euklidischen Distanz nicht normiert ist. Es hängt von der Zahl der Variablen ab und die Distanzen können größer 1 sein. Bei den Propensity Scores ist die maximale Distanz dagegen 1.

handlung, ist dies nicht der Fall. Hinzu kommt, dass die City-Block-Metrik eine ordinale Interpretation besitzt (ebenda).

18 Aus Gründen der Einfachheit der Schreibweise wird im Folgenden auf die Gewichtung mit $1/s_{ik}$ verzichtet. Daraus resultieren keine Beschränkungen.

Modifikationen

Für das statistische Matching wurden mehrere Modifikationen vorgeschlagen (***Smith*** 1997, S. 334-341), von denen hier nur einige erwähnt werden sollen.

- *Verwendung der linearen Regression* bzw. der *Diskriminanzanalyse* an Stelle der logistischen Regression, da nicht ausgeschlossen werden kann, dass die lineare Regression zu besseren Prognosewerten führt (***Smith*** 1997, S. 335).
- *Verwendung der City-Block-Metrik* an Stelle der quadrierten euklidischen Distanz. Als Vorteil der City-Block-Metrik könnte angeführt werden, dass sie eine ordinale Interpretation besitzt (***Bacher*** 1996, S. 216-217). Dem steht als möglicher Nachteil oder Vorteil gegenüber, dass im Unterschied zur (quadrierten) euklidischen Distanz große Abweichungen in einer Variablen nicht stärker gewichtet werden als viele kleine (ebenda, S. 222). Als weiteren Nachteil oder Vorteil eines der beiden Maße kann gesehen werden, dass bei der quadrierten euklidischen Distanz Variablen mit einer größeren Streuung ein geringeres Gewicht erhalten (ebenda, S. 180-185).
- *Mehrfache Verwendung eines Falles als statistischer Zwilling* (***Smith*** 1997, S. 338). Ein einmal als Zwilling identifizierter Fall wird für die weitere Zuordnung nicht gestrichen, sondern kann erneut als Zwilling verwendet werden. Gegen dieses Vorgehen spricht, dass Autokorrelationen entstehen können. Als Vorteil kann genannt werden, dass bessere Zwillinge gefunden werden, da immer alle Fälle zur Auswahl stehen.
- *Multiples Matching.* Analog zu Techniken bei der Behandlung fehlender Werte (***Rubin*** 1987; ***Rässler*** 2001, S. 71-75) wird das Verfahren mehrfach mit verschiedenen zufälligen Anordnungen gerechnet, um Verzerrungen, wie eine Unterschätzung von Varianzen und Standardfehlern, zu vermeiden.
- *Verwendung multipler Zwillinge* (***Smith*** 1997, S. 339). Jedem Datensatz wird in einem Durchlauf nicht ein Zwilling, sondern mehrere Zwillinge zugeordnet. Dadurch kann die Effizienz der Schätzer erhöht werden. Umgekehrt nimmt das Risiko von Verzerrung zu, da die Wahrscheinlichkeit der Auswahl von "schlechten" Zwillingen, die nur mehr eine geringe Ähnlichkeit zu ihren statistischen Geschwistern besitzen, steigt.

4. Umsetzung der Verfahren in SPSS-Syntaxprogramme

Das Programm für das statistische Matching ist im Anhang A1 wiedergegeben und dort ausführlich dokumentiert. Das eigentliche Matching erfolgt in der SPSS-Matrixsprache, die mit dem Befehl MATRIX aufgerufen wird. Mit der Anweisung

GET M/VARIABLES = BEHANDL IDNR PRE_1 GEFUNDEN wird die zu bearbeitende Matrix definiert. Nach der Spezifikation der Größe der Untersuchungsgruppe, der Gesamtfallzahl und des Schwellenwertes CC erfolgt die Zwillingssuche durch LOOP-Schleifen. Durch die äußere Schleife LOOP #i=1 to N1, die vor SAVE mit dem Befehl END LOOP abgeschlossen wird, werden die Behandlungsfälle abgearbeitet. Im ersten Durchlauf ist #i gleich 1, im zweiten gleich 2 usw. Vor der statistischen Zwillingssuche werden die Variablen II, DD und DDD initialisiert. In der Variablen II soll später der Index des statistischen Zwillings stehen. In DDD soll die Distanz zwischen dem Datensatz #i und dem potentiellen Zwilling #j stehen, in DD der bisher kleinste gefundene Wert. Durch die Anweisung COMPUTE DDD=DD+1 wird bewirkt, dass die Distanz DDD zu Beginn auf jeden Fall größer als DD ist. Die Suche nach dem statistischen Zwilling für einen Datensatz #i erfolgt in der Schleife LOOP #j=i2 to N2. Durch den Befehl DO IF .. wird garantiert, dass nur die Fälle einbezogen wurden, die noch nicht als statistische Zwillinge verwendet wurden. In diesem Fall ist der Wert der vierten Spalte (COL4) gleich 0. Mit COMPUTE DDD=abs(m(#i,3)-m(#j,3)) wird die absolute Differenz der Propensity Scores von #i und #j berechnet. In der anschließenden DO IF-Abfrage wird überprüft, ob die Differenz kleiner ist als die bisher kleinste berechnete Distanz. Falls dies zutrifft, ist der Datensatz #j besser als Zwilling geeignet. Sein Index wird in die Variable II geschrieben und die Distanz DDD wird als neuer Kriteriumswert definiert. Mit END LOOP wird die Zwillingssuche für den Datensatz #i abgeschlossen. In die vierte Spalte wird – sofern die Distanz DD für den potentiellen Zwilling kleiner dem Schwellenwert CC ist – die ALLBUS-Identifikationsnummer von #i geschrieben, die in der zweiten Spalte von M steht. In die erste Spalte wird für spätere Analysen die Distanz zwischen den beiden Propensity Scores eingetragen. Anschließend wird die Suche mit #i=#i+1 fortgesetzt.

Der Anwender kann das Programm für seine Zwecke weitgehend mit geringfügigen Modifikationen übernehmen. Folgende Änderungen sind notwendig:

- Der Dateiname in dem GET FILE-Befehl muss angepasst werden.
- Die Identifikationsvariable muss geändert werden. I.d.R. wird dies nicht V2 sein.
- Die Fallzahl N1 der Untersuchungsgruppe (in unserem Beispiel der Arbeitslosen), die Gesamtfallzahl N2 und der Schwellenwert CC müssen geändert werden.
- Die Variablen, die in die anschließenden statistischen Analysen einbezogen werden, müssen angepasst werden.

Sollen an Stelle der Propensity Scores Distanzen berechnet werden, sind einige Programmzeilen zu ändern (siehe Anhang A2):

- Die Variablen müssen zuvor gewichtet werden. Dies geschieht durch die Anweisungen compute zOST=ost/0.47272; compute zalter=alter/12.18747 usw.
- In die Matrix M müssen die gewichteten Variablenwerte geschrieben werden (GET M /VARIABLES= behandl IDNR pre_1 gefunden zost to zgewerk). Sie stehen in unserem Beispiel in den Spalten COL5 bis COL20.
- Die Berechnung der Distanzen macht eine weitere Schleife

  ```
  +     loop #k=5 to 20.
  +     compute ddd=ddd+(m(#i,#k)-m(#j,#k))*(m(#i,#k)-m(#j,#k)).
  +     end loop.
  ```

  erforderlich, die über die Spalten läuft, in denen die Variablenwerte stehen.

Weitere Modifikationen sind nicht erforderlich. Der Durchlauf der Programme, insbesondere jenes mit den quadrierten euklidischen Distanzen, kann mitunter etwas Zeit dauern.[19]

5. Ergebnisse

In einem ersten Durchlauf wurde mit sehr großen Schwellenwerten *c* gerechnet, um zu erreichen, dass für alle Fälle ein statistischer Zwilling gefunden wird. Dabei geht man das Risiko ein, dass auch "schlechte" Zwillinge mit deutlichen Unterschieden in den Variablen gebildet werden. Ob die Zwillingssuche erfolgreich war, kann dadurch untersucht werden, dass die Korrelationen zwischen der Behandlungsvariable und den Kontrollvariablen berechnet werden. Sie müssen 0 sein. Beide Verfahren (Propensity Scores und Distanzmodell) erfüllen diese Bedingung (siehe Tabelle 3). Alle bivariaten Korrelationen sind im Unterschied zu einer Analyse mit den Ausgangsdaten (alle Arbeitslosen und alle Erwerbstätigen) nicht signifikant. Auch die durchgeführten t-Tests (Ergebnisse hier nicht wieder gegeben) erbrachten keine signifikanten Unterschiede.

19 Auf meinem Notebook mit einem Intel Celeron Prozessor benötigte das Programm mit den quadrierten euklidischen Distanzen mehrere Minuten.

Tabelle 3: Korrelationen der Kontrollvariablen mit der Behandlungsvariablen

		Ausgangsdaten	stat. Zwillinge (Propensity Scores)	stat. Zwillinge (quadrierte euklidische Distanzen)
		BEHANDL	BEHANDL	BEHANDL
	N	1809	360	360
OST	Pearson Correlation	,204***	,000	,018
ALTER	Pearson Correlation	,117***	,053	,060
WEIBL	Pearson Correlation	,075**	-,006	-,022
VERH	Pearson Correlation	-,015	,045	,000
VERHGE	Pearson Correlation	,051*	-,039	,000
VERW	Pearson Correlation	,053*	-,014	,000
GESCHIED	Pearson Correlation	,071**	-,025	,000
LEDIG	Pearson Correlation	-,055*	-,007	,000
PRES	Pearson Correlation	-,175***	,018	-,069
SCHULB	Pearson Correlation	-,175***	-,023	-,080
KEINE	Pearson Correlation	,141***	,000	,000
RK	Pearson Correlation	-,135***	-,056	-,010
EV	Pearson Correlation	-,027	,043	,006
ANDERE	Pearson Correlation	,035	-,018	,000
GEWERK	Pearson Correlation	-,032	-,062	,000
LINKSRE	Pearson Correlation	-,027	,052	-,010

*** p < 0,1%; ** p < 1%; * p <5%

Effektberechnung. Zur Berechnung des Einflusses der Behandlungsvariablen auf die abhängige Variable stehen zwei Möglichkeiten zur Verfügung:

- Mit Hilfe des t-Tests oder eines anderen geeigneten Verfahrens wird untersucht, ob sich die beiden Gruppen in der abhängigen Variablen signifikant unterscheiden. Die anderen unabhängigen Variablen werden bei diesem Vorgehen nicht mehr berücksichtigt.
- Mit Hilfe der multiplen Regression oder eines anderen geeigneten multivariaten Verfahrens wird der Effekt der Behandlungsvariablen geschätzt. Die anderen unabhängigen Variablen gehen als Kontrollvariable in die Analyse ein.

Streng genommen würde die erste Vorgehensweise ausreichen, wenn das statistische Matching zu einer perfekten Übereinstimmung geführt hat, wenn also gelten würde: $\hat{p}_g = \hat{p}_{g^*}$, wobei $\hat{p}$ die Prognosewerte (Propensity Scores) sind. In diesem Fall besitzen die Propensity Scores die Eigenschaft von balancierten Scores (***Rässler*** 2001, S. 26-27). Diese garantieren, dass Kontroll- und Untersuchungsgruppe identische Verteilungen in den Kontrollvariablen haben. Die Bedingung $\hat{p}_g = \hat{p}_{g^*}$ ist in der Praxis nie erfüllt, so dass aus Sicherheitsgründen der Einfluss der Kovariaten kontrolliert werden sollte.

In unserem Anwendungsbeispiel führt eine multiple Regression zu folgenden Ergebnissen (Signifikanz in Klammern):

	Regr.koeff. von Arbeitslosigkeit auf Ausländerfeindlichkeit
Ausgangsdaten	0,412 (p<0,1%)
Statistische Zwillinge (Propensity Scores aus logist. Regr.)	0,290 (p<1%)
Statistische Zwillinge (quadrierte euklidische Distanzen)	0,138 (p>10%)

Die Ausgangsdaten und die mit Hilfe der Propensity Scores erzeugten Kontrollgruppe resultieren in einer signifikanten Wirkung der Arbeitslosigkeit, die bei Verwendung der (quadrierten) euklidischen Distanzen verschwindet.

Um mehr Klarheit über die Ursachen dieser Unterschiede zu erlangen, wurden zusätzlich die Propensity Scores aus der linearen Regression[20] und die City-Block-Metrik[21] als Distanzmaß verwendet. Des Weiteren wurde jede Analyse mehrfach mit unterschiedlichen zufälligen Anordnungen wiederholt (multiples Matching; siehe oben). Tabelle 4 fasst die Ergebnisse zusammen. Ihr ist zu entnehmen, dass mit Ausnahme der Verwendung von Propensity Scores aus der Logitanalyse der Arbeitslosigkeit keine signifikante Wirkung zukommt. Die kleinsten und damit insignifikantesten Koeffizienten treten bei der City-Block-Metrik auf. Der durchschnittliche Fehler 1. Art liegt bei über 50%. Die zweit kleinsten Koeffizienten ergeben

20 Die lineare Regression hat hier den Vorteil, dass sie auch negative Prognosewerte zulässt und daher Fälle mit Prognosewerten nahe 0 besser trennt als die logistische Regression. Bei einer Kausalanalyse ist dieser "Vorteil" selbstverständlich unerwünscht. Siehe dazu auch Abschnitt 4.

21 Die City-Block-Metrik führt dazu, dass viele kleine Abweichungen in den Variablen für genauso bedeutsam gehalten werden wie eine große Differenz in einer Variablen. Bei der (quadrierten) euklidischen Distanz werden die Unterschiede implizit gewichtet. Eine starke Abweichung in einer Variablen erhält mehr Gewicht als viele kleine Abweichungen. Siehe dazu auch Abschnitt 4.

sich für die quadrierten euklidischen Distanzen. Aber auch die Propensity Scores auf der Grundlage der linearen Regression führen bei einem Schwellenwert von 5% zu insignifikanten Einflüssen der Arbeitslosigkeit.

Tabelle 4: Unstandardisierte Regressionskoeffizienten der Arbeitslosigkeit auf Ausländerfeindlichkeit für unterschiedliche zufällige Anordnungen

	1. Versuch	2. Versuch	3. Versuch	4. Versuch	5. Versuch	Durch-schnitt (a)
Ausgangsdaten	0,412***	0,412***	0,412***	0,412***	0,412***	0,412***
Statistische Zwillinge (Propensity Scores aus logist. Regression)	0,290**	0,284**	0,295**	0,289**	0,301**	0,292**
Statistische Zwillinge (quadrierte euklidi-sche Distanzen)	0,138	0,102	0,106	0,128	0,105	0,116
Statistische Zwillinge (Propensity Scores aus linearer Regression)	0,187	0,198	0,177	0,201	0,177	0,188
Statistische Zwillinge (City-Block-Distanzen)	0,070	0,081	0,059	0,055	0,038	0,061

*** p < 0,1%; ** p < 1%

Offensichtlich weisen die Propensity Scores aus der Logitanalyse Eigentümlichkeiten auf, die zu abweichenden Ergebnissen führen. Um diese zu erkunden, wurde in einem nächsten Schritt der Schwellenwert *c* (siehe Abschnitt 4) systematisch reduziert. An die statistischen Zwillinge wurden also immer strengere Anforderungen bezüglich der Ähnlichkeit gestellt. Bis zu einem Schwellenwert von 0,001 wird nach wie vor ein signifikanter Einfluss der Arbeitslosigkeit berechnet. Bei 0,0001 ergibt sich ein insignifikanter Regressionskoeffizient von 0,175 (p>10%; verbleibendes n=128).

Das Ergebnis lässt sich gut erklären: In unserem Beispiel treten große Abweichungen der Propensity Scores bei höheren Propensity Scores, also bei einem hohen Arbeitslosigkeitsrisiko, auf. Hier zeigt sich, dass der Propensity Score der Arbeitslosen i.d.R. höher ist als jener der Erwerbstätigen. Es gilt also: $\hat{p}_{g*} < \hat{p}_g$, wenn $\hat{p}_g$ hoch

ist. Ein hoher Propensity Score geht nun aber mit einer höheren Ausländerfeindlichkeit einher, d.h., umso höher $\hat{p}$ ist, desto größer ist die Ausländerfeindlichkeit. Da Arbeitslose höhere Propensity Scores haben, sind sie auch ausländerfeindlicher. Es liegt also ein systematischer Bias vor, der zu einer signifikanten Wirkung der Arbeitslosigkeit führt. Dieser Bias verschwindet, wenn ein Schwellenwert eingeführt wird.

6. Zusammenfassung und Diskussion

Aufgabe des statistischen Matching ist das Auffinden von statistischen Zwillingen. Statistische Zwillinge sind dadurch gekennzeichnet, dass sie sich untereinander in ausgewählten Merkmalen nicht unterscheiden. Sie können für ein breites Spektrum von Anwendungen eingesetzt werden. In der Praxis werden sie – abgesehen von der Behandlung fehlender Werte – kaum eingesetzt. Eine Ursache hierfür sind vermutlich fehlende Programmmodule in Standardstatistikprogrammen, wie SPSS. Das Hauptziel des Beitrages war daher darzustellen, wie statistische Zwillinge mit Hilfe eines SPSS-Syntaxprogrammes berechnet werden können. Syntaxprogramme für zwei Methoden wurden behandelt, nämlich für Propensity Scores und Distanzfunktionen. Das Vorgehen und die Berechnung wurden anhand eines Forschungsbeispiels aus dem ALLBUS 1996 erörtert.

Die Ergebnisse zeigen, dass beim statistischen Matching – ausgenommen den Ergebnissen bei Verwendung einer logistischen Regression – der in einer multivariaten Analyse ohne Matching ermittelte signifikante Zusammenhang zwischen Arbeitslosigkeit und Ausländerfeindlichkeit verschwindet. Die abweichenden Ergebnisse für die logistischen Propensity Scores erklären sich dadurch, dass ein systematischer Zusammenhang besteht zwischen Propensity Scores, Behandlungsstatus und Ausländerfeindlichkeit. Es empfiehlt sich daher, immer zu untersuchen, ob das statistische Matching zu systematischen Fehlern führt. Eine Betrachtung von Mittelwertunterschieden in den Kovariaten zwischen den beiden Gruppen oder von Korrelationen der Kovariaten mit den beiden Gruppen muss nicht ausreichend sein.

Das Beispiel macht deutlich, dass der Einsatz von statistischen Matchingverfahren fruchtbar und notwendig sein kann, wenn eine heterogene, im Vergleich zur Untersuchungsgruppe relativ große Vergleichsgruppe vorliegt. Heterogen bedeutet dabei, dass sich die heterogene Vergleichsgruppe aus der eigentlichen, häufig aber unbekannten Vergleichsgruppe, die sich in den Kontrollvariablen nicht von der Untersuchungsgruppe unterscheidet, plus weiteren Subpopulationen zusammensetzt. Die Subpopulationen unterscheiden sich sowohl hinsichtlich der Verteilung in den Kontrollvariablen als auch hinsichtlich des Zusammenhangs zwischen Kontrollvariablen

und abhängigen Variablen von der eigentlichen mit der Untersuchungsgruppe statistisch weitgehend identischen Vergleichsgruppe.

Das statistische Matching kann aber auch noch für zahlreiche weitere Anwendungen eingesetzt werden. In Zukunft wird das statistische Matching meiner Einschätzung nach an Bedeutung gewinnen, da in einem größeren Umfang als bisher Registerdaten zur Verfügung stehen werden, aus denen sich Kontrollgruppen bilden lassen.

Erwähnt werden sollen aber auch derzeit noch ungelöste Probleme und offene Fragen.[22] Bei der Analyse wurde der Empfehlung gefolgt, alle Variablen für das statistische Matching zu verwenden, die einen Einfluss auf die abhängige Variable haben können. Dieser Empfehlung lässt sich nur bei der Verwendung von Propensity Scores uneingeschränkt zustimmen. Irrelevante Variable haben hier keinen bzw. nur einen geringen Einfluss auf die Ergebnisse. Im Unterschied dazu können irrelevante Variable bei Distanzfunktionen zu einer entscheidenden Verschlechterung der Ergebnisse führen, wie wir an anderer Stelle durch Modellrechnungen für das empirische Re-Identifikationsrisiko einer Registerdatei durch eine Umfrage nachweisen konnten (***Bacher***, ***Brand*** und ***Bender*** 2002). Auf der anderen Seite führen Distanzfunktionen – wie in unserem Beispiel – möglicherweise aber zu einem besseren Matching und wären deshalb zu bevorzugen. Sie sind auch flexibler. Sollte sich z.B. zeigen, dass sich die beiden Gruppen auch nach dem Matching in einer Variablen unterscheiden, so kann der Variablen bei einer Wiederholung des Matching ein höheres Gewicht gegeben werden[23].

Gegen das statistische Matching lässt sich schließlich anführen, dass das Ziel der Sozialforschung das Auffinden komplexer Zusammenhangsmuster ist und nicht das Auffinden eines Einzelzusammenhanges. Besser geeignet – so der Einwand – sind hierfür Pfad- bzw. Strukturgleichungsmodelle. Diese Argumentation ist teilweise gerechtfertigt. Es gibt nicht d a s Patentverfahren, das sich für alle Anwendungen eignet. Dies gilt aber vice versa auch für Pfad- und Strukturgleichungsmodelle.

22 Zu beachten ist hierbei, dass sich der Autor noch nicht vollständig in die sehr umfangreiche Literatur eingelesen hat. Die Aufstellung ist daher selektiv.

23 Bei der Verwendung der Propensity Scores kann eine Gewichtung dadurch erreicht werden, dass zusätzlich die quadrierten Variablenwerte in die Berechnung einbezogen werden. Bei nominalen und dichotomen Variablen ist dieses Vorgehen aber nicht möglich.

Literatur

Aisenbrey, S. 2000:
Optimal Matching Analyse. Anwendungen in den Sozialwissenschaften. Opladen: Leske und Budrich.

Althauser, R.; und ***Rubin, D***. 1970:
The Computerized Construction of a Matched Sample. American Journal of Sociology, 76, 2, pp. 325-346.

Althauser, R.; und ***Rubin, D***. 1971:
Measurement Error and Regression to the Mean in Matched Samples. Social Forces, 50, pp. 206-214.

Bacher, J. 1996:
Clusteranalyse. 2. Auflage. München-Wien: Oldenburg.

Bacher, J.; Brand, R.; und ***Bender,*** S. 2002:
Re-Identifying Register Data by Survey Data Using Cluster Analysis: An empirical Study. erscheint in: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

Bender, S.; Brand, R. und ***Bacher, J.*** 2001:
Re-identifying register data by survey data: An empirical study. Statistical Journal of the United Nations Economic Commission for Europe, Vol. 18, Number 4, Special Issue: Data Confidentially, pp. 373-381.

Blasius, J. 2001:
Korrespondenzanalyse. München-Wien: Oldenburg.

Blasius, J. und ***Thiessen, V***. 2001:
Methodological artefacts in measures of political efficacy and trust: a multiple Correspondence Analysis. Political Analysis, Vol 9, pp. 1-20.

Brand, R. 2000:
Anonymität von Betriebsdaten. Nürnberg: Beiträge zur Arbeitsmarkt- und Berufsforschung (BeitrAB) 237.

Chapin, F. S. 1974:
Experimental Designs in Sociological Research. Revised Edition. Connecticut: Greenwood Press.

Fröhlich, M. 2002:
Programme Evaluation with Multiple Treatments. St. Gallen: Discussion paper no 2002-17, Department of Economics.

Gerfin, M.; und ***Lechner, M.*** 2000:
Microeconometric Evaluation of the Active Labour Market Policy in Switzerland. Bonn: IZA, Discussion Paper No. 154.

Holm, K. 2001:
ALMO-Data-Mining. Ein Standard-Auswertungssystem. Linz: Eigenverlag (Handbuch).

Lechner, M. 1999:
Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption. Bonn: IZA, Discussion Paper No. 91.

Little, R. J. A.; und ***Rubin, D. B.*** 1987:
Statistical Analysis with Missing Data. New York: John Wiley and Sons. (wurde 2002 als Wiley-Classic neu aufgelegt)

Rässler, S. 2001:
Alternative Approaches to Statistical Matching with an Application to Media Data. Nürnberg: Habilitationsschrift. (Die Habilitationsschrift liegt in der Zwischenzeit auch als Buchpublikation vor: Rässler, S., 2002: Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches, Lecture Notes in Statistics, 168. New York: Springer.)

Rubin, D. B. 1987:
Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Smith, H. L. 1997:
Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. Sociological Methodology 1997, Vol. 27, pp. 325-353.

United Nations (Ed.) 2001:
Statistical Journal of the United Nations Economic Commission for Europe, Vol. 18, Number 4, Special Issue: Data Confidentially.

Anhang A1: SPSS-Syntaxprogramm zur Suche von statistischen Zwillingen mit Hilfe von Propensity Scores

```
*================================================================.
* Analysedatei einlesen. Diese muss beinhalten:
*  (a) die Kontrollvariablen, die Behandlungsvariablen
*      und die abhängige Variable.
*  (c) die Propensity Scores (Prognosewerte)
*  (d) Zufallszahlen für die zufällige Anordnung der Fälle.
* Die Propensity Scores stehen in der Datei PP1.SAV in der
* Variablen PRE_1, die Zufallszahlen in den Variablen H1 bis H5.
* Die Propensity Scores wurden berechnet mit:
* logistic regr var= .....
*      /meth=enter
*      /save=pred.
* Die Zufallszahlen wurden erzeugt mit:
* compute h1=rv.uniform(0,10000).
* compute h2=rv.uniform(0,10000).
* compute h3=rv.uniform(0,10000).
* compute h4=rv.uniform(0,10000).
* compute h5=rv.uniform(0,10000).
* Die große Spannbreite von 0 bis 100000 garantiert, dass ein Wert
* nicht mehrmals auftritt.
* Die Behandlungsvariable muss 0 (=Keine Behandlung; in unserem
* Beispiel nicht arbeitslos) 1 (=Behandlung; arbeitslos)
* kodiert sein.
* Alle Fälle mit fehlenden Werten in den Propensity Scores müssen
* von der Analyse ausgeschlossen werden.
*================================================================.
get file="c:\texte\datenanalyse\pp1.sav".
*================================================================.
* Datei nach Behandlungsstatus (absteigend, also 1 zuerst und
* dann 0) und Zufallsvariable (aufsteigend) sortieren.
sort cases by behandl (d) h1 (a).
```

```
execute.
*================================================================.
* Die Hilfsvariable GEFUNDEN erzeugen.
compute gefunden=0.
* Identifikationsvariable IDNR definieren. Sie
* enthält die fortlaufende Befragtennummer, die in
* V2 steht.
compute idnr=v2.
*Zahl der maximalen Schleifen erhöhen.
SET MXLOOP=300000.
* MATRIX M definieren.
MATRIX.
GET M /VARIABLES= behandl idnr pre_1 gefunden.
*================================================================.
* Die Variablen BEHANDL, IDNR usw. werden intern
* mit COL1, COL2 usw. bezeichnet.
*
* In n1 die Zahl der Arbeitslosen eingeben,
* in n2 die Gesamtfallzahl,
* in cc den Schwellenwert für die Ähnlichkeit der Zwillinge.
* Für den Schwellenwert wird zunächst ein sehr hoher Wert
* verwendet, um zu gewährleisten, dass für alle Fälle
* ein Zwilling gefunden wird.
*================================================================.
compute n1=180.
compute i2=n1+1.
compute n2=1809.
compute cc=0.5.
loop #i=1 to n1.
+  compute ii=0.
+  compute dd=9999.
+  compute ddd=dd+1.
+   loop #j=i2 to n2.
+     do if (m(#j,4) eq 0).
+         compute ddd=abs(m(#i,3)-m(#j,3)).
+     end if.
+     do if (ddd < dd).
+         compute ii=#j.
+         compute dd=ddd.
+      end if.
+   end loop.
+   do if (dd < cc).
+       compute m(#i,4)=m(#i,2).
```

```
+       compute m(ii,4)=m(#i,2).
+       compute m(#i,1)=m(#i,3)-m(ii,3).
+       compute m(ii,1)=m(#i,3)-m(ii,3).
+   end if.
end loop.
save M/outfile=*.
END MATRIX.
select if (col4 > 0).
sort cases by col4.
execute.
*=============================================================.
*Datenmatrix visuell prüfen!
*Die statistischen Zwillinge stehen untereinander.
*Ihre Prognosewerte müssen relativ ähnlich sein.
*Fälle nach der Identifikationsnummer sortieren.
sort case by col2.
compute idnr=col2.
*Matrix zwischenspeichern.
save outfile="c:\texte\datenanalyse\mm.sav".
*Ausgangsdatei laden, nach Befragtennummer
*sortieren und wieder abspeichern.
get file="c:\texte\datenanalyse\pp1.sav".
compute idnr=v2.
sort cases by idnr.
save outfile="c:\texte\datenanalyse\pp1.sav".
execute.
*Analysevariablen an die Datei MM anhängen.
match files
   file="c:\texte\datenanalyse\mm.sav"
   /table="c:\texte\datenanalyse\pp1.sav"
   /by idnr
   /map.
execute.
*=============================================================.
*Mit T-TEST und CORR prüfen, ob sich Untersuchungsgruppe
*und Kontrollgruppe in den Kontrollvariablen unterscheiden.
*Es sollten keine signifikanten Unterschiede auftreten.
t-test groups=behandl (0,1)/var =behandl ost alter weibl verh to ledig
       schulb pres rk to keine linksre gewerk.
corr ost alter weibl verh to ledig  schulb pres rk to keine
       linksre gewerk with behandl.
*=============================================================.
* War das Matching nicht erfolgreich, dann sollte der Vorgang mit
```

```
* einem anderen Schwellenwert wiederholt werden.
* Nicht erfolgreich bedeutet:
* Eine oder mehrere Kontrollvariablen korrelieren signifikant
* mit der Behandlungsvariablen.
* Der t-Test erbrachte signifikante Unterschiede.
*=============================================================.
*Effekt der Behandlungsvariable auf die abhängige Variable ermitteln.
t-test groups=behandl (0,1)/var =ausl1.
regr var=ausl1 behandl ost alter weibl verhge to ledig  schulb pres
       rk to andere linksre gewerk
      /dep=ausl1
      /meth=enter.
Anhang A2: SPSS-Syntaxprogramm zur Suche von statistischen Zwillingen mit
Hilfe von quadrierten euklidischen Distanzen
*Kommentare siehe Anhang A1.
get file="c:\texte\datenanalyse\pp1.sav" .
*Datei nach Behandlungsstatus und Zufallsvariable sortieren.
sort cases by behandl (d) h5 (a).
execute.
*Hilfsvariable GEFUNDEN berechnen.
compute gefunden=0.
*Standardabweichungen für die Variablen der Untersuchungsgruppe
*berechnen und ausdrucken.
temp.
select if (behandl=1).
desc var=ost alter weibl verh to ledig  schulb pres rk to keine linksre
gewerk.
*---Programm nur bis zu dieser Stelle laufen lassen!
*Ausgedruckte Standardabweichungen für die Gewichtung verwenden.
compute zOST=ost/0.47272.
compute zalter=alter/12.18747.
compute zWEIBL=weibl/0.50112.
compute zVERH=0.7071*verh/0.48618 .
compute zVERHG=0.7071*verhge/0.19387.
compute zVERW=0.7071*verw/0.19387.
compute zGESCH=0.7071*geschied/0.32192.
compute zLEDIG=0.7071*ledig/0.38802 .
compute zSCHULB=schulb/0.72849.
compute zPRES=pres/20.08525.
compute zRK=0.7071*rk/0.27716 .
compute zEV=0.7071*ev/0.45954.
compute zANDER=0.7071*andere/0.14782.
compute zKEINE=0.7071*keine/0.49237.
```

```
compute zLINKSRE=linksre/1.84444.
compute zGEWERK=gewerk/0.38802.
execute.
*Kontrollausgabe:
*Standardabweichungen von Zost und Zgewerk müssen
*für Behandl=1 gleich 1 (quantitative Variablen) oder
*gleich 0,7071 (Dummies) sein.
means tabels=zost to zgewerk by behandl.
compute idnr=v2.
SET MXLOOP=300000.
MATRIX.
GET M /VARIABLES= behandl v2 pre_1 gefunden zost to zgewerk.
compute n1=180.
compute i2=n1+1.
compute n2=1809.
compute cc=99.
loop #i=1 to n1.
+  compute ii=0.
+  compute dd=9999.
+   loop #j=i2 to n2.
+   compute ddd=0.
+     do if (m(#j,4) eq 0).
+         loop #k=5 to 20.
+              compute ddd=ddd+(m(#i,#k)-m(#j,#k))*(m(#i,#k)-m(#j,#k)).
+         end loop.
+     do if (ddd < dd).
+         compute ii=#j.
+         compute dd=ddd.
+      end if.
+      end if.
+   end loop.
+   do if (dd < cc).
+     compute m(#i,4)=m(#i,2).
+     compute m(ii,4)=m(#i,2).
+     compute m(#i,1)=m(#i,3)-m(ii,3).
+     compute m(ii,1)=m(#i,3)-m(ii,3).
+     compute m(#i,3)=dd.
+     compute m(ii,3)=dd.
+   end if.
end loop.
save M/outfile=*.
END MATRIX.
```

Rest analog zum Programm in Anhang A1.